DIVISION OF THE HUMANITIES AND SOCIAL SCIENCES
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA 91125

# SELECTION BIAS IN LINEAR REGRESSION, LOGIT AND PROBIT MODELS

Jeffrey A. Dubin
California Institute of Technology

Douglas Rivers
University of California, Los Angeles


SOCIAL SCIENCE WORKING PAPER 698

April 1989

# SELECTION BIAS IN LINEAR REGRESSION, LOGIT AND PROBIT MODELS

Jeffrey A. Dubin and Douglas Rivers

**Abstract**

Missing data are common in observational studies due to self-selection of subjects. Missing data can bias estimates of linear regression and related models. The nature of selection bias and econometric methods for correcting it are described. The econometric approach relies upon a specification of the selection mechanism. We extend this approach to binary logit and probit models and provide a simple test for selection bias in these models. An analysis of candidate preference in the 1984 U.S. presidential election illustrates the technique.

# SELECTION BIAS IN LINEAR REGRESSION, LOGIT AND PROBIT MODELS

Jeffrey A. Dubin and Douglas Rivers

## 1. Introduction

Most empirical work in the social sciences is based on observational data which is incomplete. Often data are missing for reasons other than that the investigator (or other collector of the data) did not record certain measurements. A much more common cause of missing data is that the subjects themselves act in a way that makes it impossible to obtain measurements on certain variables. For example, in political surveys we do not have data on how some respondents voted for the simple reason that some respondents *chose* not to vote. Restricting data analysis to the sample of voters leaves us with a *self-selected* sample. If our interest is in the relationship between demographic characteristics and political preferences in the population as a whole, the subsample of non-missing observations is likely to produce misleading conclusions.

In the voting example above, one solution would be to ask non-voters which candidate they would have voted for, but this "solution" is not very practical for secondary analysts who lack control over data collection. In other situations, it is hard to envision how the missing data could be collected even if we had abundant resources. For example, in analyzing the relationship between schooling and earnings, we only have earnings data for those who are employed. Labor force participation is voluntary. Some people choose not to work, others are unable to find work they considerable acceptable. The employed sample is unlikely to be a random subset of the entire population and there is no reliable way to impute earnings to those who are unemployed.

In recent years there has been a good deal of work devoted to missing data problems. (The book by Little and Rubin (1987) is a good summary; see also their paper in this volume for an alternative approach to handling missing data.) The method developed by Heckman (1979) for correcting for selectivity bias in linear regression models with normal errors has found many applications in econometrics and is now a standard tool for empirical workers. Little, however, is known about the treatment of missing data in probit and logit models. These models have attained considerable popularity in the social sciences for analyzing discrete choice and other qualitative data. Unfortunately, there is no simple analog to the Heckman method for discrete choice models, even though the same basic

conceptual framework carries over in a natural way. In this paper, we adapt the Heckman framework to logit and probit models and discuss various methods of estimation in this context.

To provide background material for readers who may be unfamiliar with the standard econometric approach to selectivity, a brief exposition of selection bias in linear regression models is presented in section 2. We restrict our attention to cases where only observations on the dependent variable are missing. The simplest case is the well-known Tobit model of Tobin (1958) where the censoring is governed by the value of the dependent variable itself. A simple geometric argument makes the nature of the bias apparent and the maximum likelihood estimator is very simple to develop in this context. We then consider Heckman's (1979) adaptation of the Tobit model to situations where there is a separate mechanism governing the censoring.

In sections 3 through 5, we adapt the Heckman setup to probit and logit analysis with selectivity. Analogous to the Heckman method, there is both a two-step estimator and a maximum likelihood estimator. The computational advantages of two step estimation are less here than in Heckman's case, as they still require specialized software. We also propose a simple score test for selection bias that does not require computation of full model. Section 6 contains estimates of a voting model that have been corrected for selection bias. All but the simplest derivations have been placed in the Appendix.

## 2. Selection Bias in Linear Regression Models

In this section we briefly review the symptoms and treatment of selection bias in linear regression models. In this case, selection bias turns out to be a garden variety specification error similar to omitting a variable. The obvious solution—of including the omitted variable—is an effective cure. The linear regression model is a convenient starting place for the subsequent development.

Our primary interest concerns a linear regression model of the form:

$$y_i = \beta' x_i + u_i. \tag{2.1}$$

Equation (2.1) is a "structural model" that is intended to represent some behavioral process. In econometric applications, (2.1) might arise from an optimization problem. For example, $y_i$ might denote the quantity consumed of some good and the vector $x_i$ would include the prices of various goods and characteristics of the consumer (including income).

Equation (2.1) could be obtained by specifying a "representative" utility function for each consumer (depending upon the quantity consumed of each good and the consumer's demographic characteristics). The quantity consumed, $y_i$, is assumed to maximize the consumer's utility subject to a budget constraint. At some point in the derivation, the error term $u_i$ is introduced to capture unmeasured variables in the utility function or, perhaps, errors of optimization.

The key point is that equation (2.1) is assumed to hold independent of how any data might be collected. It is this aspect of the econometric approach that often causes statisticians difficulty. The regression in (2.1) is not an empirical relation, but a theoretical one. At the outset we are willing to commit to a specification of how $y_i$ is generated that is derived from an economic or other social scientific theory. The purpose of estimation is not to learn what process generates the observed variables—as this is taken to be known in advance of any data analysis—but to learn the parameters of this process (such as price elasticities).

It should be obvious that (2.1) alone does not determine the distribution of the observed variables $(x_i, y_i)$. This will depend on two things: the distribution of the errors and how the data were collected. We discuss each in turn.

In most applications, the error term $u_i$ is introduced because the theory, as represented by the rest of (2.1), is not completely adequate. One should be reluctant to make too many assumptions about the errors which, admittedly, represent theoretical ignorance. But, to the degree that we have confidence in our theory, observations with large errors are unusual because they are not accounted for by the model. In this sense, the errors represent failures of the model. Being realists, we are willing to tolerate such failures so long as they have no systematic pattern. The customary assumptions are that nothing systematic has been omitted from the model ($x_i$ and $u_i$ are independent) and that, on average, the model is correct (the mean of $u_i$ is zero). Again, these assumptions appear to be largely a theoretical matter. If one believes the theory, then one should be willing to make the necessary assumptions.

Data collection is an altogether different matter. One can believe the theory implied by (2.1) in its entirety and yet not expect a sample to yield regression estimates resembling (2.1). The sampling procedure may be such that it over or under represents specific types of individuals. This causes no serious problems if the sampling fractions are purely a function of the explanatory variables. Nor is it a problem when the sampling fractions are

independent of the errors. The source of the problem is sample selection *related* to the errors. When this happens, the assumed theoretical model fails in a *systematic* way: the errors occuring in the sample no longer have a zero mean because the sampling procedure has picked out observations which are, in terms of the theory, "unusual."

### The Tobit Model

The simplest case of selection bias arises when certain observations on the dependent variable $y_i$ have been "censored." In a classic paper, Tobin (1958) analyzed automobile purchases. In this application, $y_i$ denotes the amount a household would like to spend on new cars. If the least expensive car costs $c$, households whose desired level of automotive expenditures is less than $c$ will be unable to transact. In this case, we would not observe the amount $y_i$ that they would like to spend. Their actual expenditures would be zero, but this would not be indicative of their desired expenditures which the regression is intended to explain. For any given level of $x_i$, the sample would overrepresent those households with large positive errors.

It might be tempting in this situation to go ahead and regress $y_i$ on $x_i$ and a constant using only those households who purchased cars. In Figure 1, solid dots indicate households for which $y_i \geq c$; these are households whose desired level of expenditures was sufficiently high that a transaction occurred. Empty circles indicate households with zero expenditures, i.e. those for which $y_i < c$. It is apparent from Figure 1 that use of the truncated sample can lead to severe bias. The estimated regression line, indicated by a dashed line, is less steep that the true regression line, indicated by a solid line.

How general is this result? A slightly more formal treatment is instructive. For simplicity, suppose (2.1) contains a single regressor and a constant term, as in Figure 1:

$$y_i = \alpha + \beta x_i + u_i. \tag{2.2}$$

What happens when we apply least squares to (2.2) omitting those observations for which $y_i < c$? Let $\hat{\beta}$ denote the least squares estimator of $\beta$ based on those observations satisfying the sample selection rule $y_i \geq c$. We will analyze the probability limit of $\hat{\beta}$ and will show that under rather general conditions that $\hat{\beta}$ is attenuated, i.e. $|E\hat{\beta}| < |\beta|$. That is, the least squares estimator based on the truncated sample will be attenuated; it will tend to underestimate the true impact of $x_i$ on $y_i$.

The relationship between $x_i$ and $y_i$ in the sample will reflect the impact of conditioning on the sample selection rule $y_i \geq c$. The basic idea is that the relation between $x_i$ and $y_i$ in the sample takes the form:

$$\begin{aligned} E(y_i|x_i, y_i \geq c) &= \alpha + \beta x_i + E(u_i|x_i, y_i \geq c) \\ &= \alpha + \beta x_i + E(u_i|u_i \geq c - \alpha - \beta x_i, x_i) \end{aligned} \tag{2.3}$$

The last term in (2.3) varies from one observation to another depending on the value of $x_i$. To simplify the notation, define:

$$\xi_i = E(u_i|u_i \geq c - \alpha - \beta x_i, x_i) \tag{2.4}$$

Letting $\tilde{u}_i = u_i - \xi_i$, (2.3) can be rewritten as:

$$y_i = \alpha + \beta x_i + \xi_i + \tilde{u}_i. \tag{2.5}$$

Equation (2.5) is in the form of a regression equation with a constant term and two regressors— $x_i$ and $\xi_i$.

Equation (2.5) provides the basis for a consistent estimation method in the presence of censoring. If the additional "regressor" $\xi_i$ were available, ordinary least squares could be applied to (2.5) to obtain estimates of $\alpha$ and $\beta$. For this regression to be consistent, it is necessary for the errors *in the subsample of uncensored observations* to have a mean of zero and to be uncorrelated with the regressors. The subsample of uncensored observations are those for which $y_i \geq c$. Thus, the relevant condition to assure consistency of the regression is that $\tilde{u}_i$ have a mean of zero and be uncorrelated with the regressors *conditional* upon $y_i \geq c$.

To see that the expectation of $\tilde{u}_i$ in the sample is zero, we take the expectation of $\tilde{u}_i$ conditional on the sample selection rule $y_i \geq c$:

$$\begin{aligned} E(\tilde{u}_i|y_i \geq c) &= E(u_i - \xi_i|y_i \geq c) \\ &= E\left(E(u_i|y_i \geq c, x_i) - \xi_i|y_i \geq c\right) \\ &= 0 \end{aligned} \tag{2.6}$$

by the law of iterated expectations[1] and equation (2.5). A similar argument shows that $u_i$ is uncorrelated with $x_i$ and $\xi_i$ in the sample:

$$E(\tilde{u}_i|x_i, \xi_i) = 0 \tag{2.7}$$

If observations on $\xi_i$ were available, then least squares could be applied to equation (2.5) to obtain unbiased and consistent estimates of $\alpha$ and $\beta$.

Least squares applied to the truncated sample is inconsistent because a variable—$\xi_i$—is omitted from the estimating equation. Heckman (1979) observed that the direction of the bias could be found by applying the standard omitted variables formula (Theil, 1957; Griliches, 1957). We supply the details below.

Estimating equation (2.2) amounts to estimating a misspecified version of equation (2.5). The omitted variables formula can be used to analyze the nature and direction of the resulting biases. The omitted variables formula states that, aside from sampling variation, the estimated coefficient of a variable in a regression with an omitted variable equals the true coefficient of that variable plus the coefficient of the omitted variable times the coefficient of the included variable in an "auxilliary regression" of the omitted variable on the included variables. In the present context the coefficient of the omitted variable $\xi_i$ is equal to one, so the usual specification bias formula reduces to:

$$\operatorname*{plim}_{n\to\infty} \hat{\beta} = \beta + \pi, \tag{2.8}$$

where $\pi$ is the coefficient of $x_i$ if $\xi_i$ were regressed on $x_i$ and a constant, i.e.:

$$\pi = \frac{\mathrm{Cov}(x_i, \xi_i | y_i \geq c)}{\mathrm{Var}(x_i | y_i \geq c)} \tag{2.9}$$

The direction of the bias depends upon the sign of $\pi$. In Appendix A, we show that the sign of $\pi$ is the opposite of that of $\beta$:

$$\begin{matrix} < \\ \pi = 0 \\ > \end{matrix} \quad \text{if} \quad \begin{matrix} > \\ \beta = 0 \\ < \end{matrix} \tag{2.10}$$

[1] The law of iterated expectations states that if $X$ is a random variable and $A$ and $B$ are events, then $E(X|A) = E[E(X|A \cap B)|A]$. See, for example, Billingsley, 1987, Theorem 34.4.

From (2.10), it follows that plim $\hat{\beta} < \beta$ if $\beta > 0$ and plim $\hat{\beta} > \beta$ if $\beta < 0$. Thus, selection biases the estimated coefficient towards zero. [2]

Since the results above indicate that a direct application of least squares is unsuitable to a truncated sample, alternative estimation procedures must be sought. Tobin (1958), in his classic paper, suggested assuming a normal distribution for $u_i$ and estimating $\alpha$ and $\beta$ by maximum likelihood. We assume that $u_i$ has a $N(0, \sigma^2)$ distribution and either that the explanatory variables in (2.1) are fixed or the analysis is conditional upon the $x$'s.[3] For censored observations ($y_i < c$), the likelihood is given by:

$$\Pr(y_i < c|x_i) = \Pr(u_i < c - \beta' x_i) = \Phi\left(\frac{c - \beta' x_i}{\sigma}\right) \tag{2.11}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function (cdf) of a standard normal random variable. For the uncensored observations ($y_i \geq c$), the distribution of $y_i$ is the same as that of $u_i$ except for its expectation (since the Jacobian of the transformation from $u_i$ to $y_i$ is unity), and is given by the density:

$$f_Y(y_i) = f_u(y_i - \beta' x_i) = \frac{1}{\sigma}\phi\left(\frac{y_i - \beta' x_i}{\sigma}\right) \tag{2.12}$$

Let $d_i$ be a dummy variable indicating whether an observation was censored ($d_i = 1$ if $y_i < c$) or not ($d_i = 1$ otherwise). Combining (2.11) and (2.12), we obtain the log-likelihood function:

$$L(\beta, \sigma) = \sum_{i=1}^{n} d_i \log \Phi\left(\frac{c - \beta' x_i}{\sigma}\right) + (1 - d_i) \log \frac{1}{\sigma}\phi\left(\frac{y_i - \beta' x_i}{\sigma}\right) \tag{2.13}$$

The ML estimates $\tilde{\beta}$ and $\tilde{\sigma}$ are obtained by maximizing (2.13) with respect to $\beta$ and $\sigma$. We will not go into the details here, except to mention that computer software is available for this problem.[4]

---

[2] Note, however, that plim $\hat{\beta} = \beta$ if $\beta = 0$, so the usual $t$-test of the hypothesis $\beta = 0$ is consistent.

[3] If the marginal distribution of the $x$'s does not involve either $\beta$ or $\sigma^2$, the conditional and full ML estimates will coincide, as for the normal linear regression model.

[4] Estimators for the models in this article have been implemented in version 2.0 of Statistical Software Tools (Dubin and Rivers, 1989).

It is also possible to estimate the Tobit model using a two-step procedure. The first stage is a probit analysis and the second stage is a linear regression. To simplify the notation, suppose the model includes only a constant and a single regressor. In the first stage, define a dummy variable $d_i$ which equals zero if the observation is censored ($y_i < c$) and equals one otherwise. Let $\alpha^* = (\alpha - c)/\sigma$ and $\beta^* = \beta/\sigma$. Since,

$$\Pr(d_i = 1|x_i) = \Phi(\alpha^* + \beta^* x_i), \tag{2.14}$$

a probit analysis with $d_i$ as the dependent variable and $x_i$ and a constant as independent variables gives consistent estimates of $\alpha^*$ and $\beta^*$. Denote these estimates by $\hat{\alpha}^*$ and $\hat{\beta}^*$, respectively. Under the assumption of normality, the mean of $u_i$ for a censored observation is given by:

$$\xi_i = \sigma\lambda(\alpha^* + \beta^* x_i) \equiv \sigma\lambda_i. \tag{2.15}$$

where:

$$\lambda(t) = \frac{\phi(t)}{1 - \Phi(t)} \tag{2.16}$$

is the reciprocal of it Mills' ratio, also called the hazard rate. (A similar formula holds when $u_{1i}$ has a non-normal distribution; see the Appendix for further discussion.) An estimate of $\lambda_i$ is available from the first stage of the procedure:

$$\hat{\lambda}_i = \phi(\hat{\alpha}^* + \hat{\beta}^* x_i). \tag{2.17}$$

Substituting (2.15) into (2.5) yields and replacing $\lambda_i$ by $\hat{\lambda}_i$ yields:

$$y_i = \alpha + \beta x_i + \sigma\hat{\lambda}_i + \tilde{u}_i + \sigma(\lambda_i - \hat{\lambda}_i). \tag{2.18}$$

In the second stage, $\alpha$, $\beta$, and $\sigma$ can be estimated by applying least squares to (2.18). There are two sources of inefficiency to this procedure. The first stage estimates of $\lambda_i$ do not fully exploit the sample information (by neglecting the values of $y_i$ for uncensored observations). Second, the errors in (2.18) are heteroscedastic (whether or not $\lambda_i$ is estimated), so that ordinary least squares is inefficient. In principle, the efficiency of the two-step procedure could be improved by using weighted least squares, but in practice it is simpler to resort to the ML estimator which is fully efficient.

**Heckman's Selectivity Model**

The simple Tobit model is only applicable when the sample selection rule depends solely on the value of the dependent variable. In other situations, the selection criterion may be correlated with the dependent variable, but other factors also affect whether a value is censored. The approach to selection bias that we pursue here involves a further specification of the sample selection mechanism. This requires a slight shift in notation. Rewrite the structural equation (2.1) that we want to estimate as:

$$y_{1i} = \beta_1' x_{1i} + u_{1i} \tag{2.19}$$

where $y_{1i}$ is only partially observable, i.e. some observations on $y_{1i}$ are censored. In this context, equation (2.19) is sometimes called the *outcome equation* to distinguish it from the *selection equation* defined below. Let $y_{2i}$ be a dummy variable indicating whether $y_{1i}$ is observed ($y_{2i} = 1$) or not ($y_{2i} = 0$). It is necessary to specify how $y_{2i}$ is determined. Since $y_{2i}$ is dichotomous, a regression model would be ill-suited for this purpose. Instead we introduce an auxilliary latent variable $y_{2i}^*$ which is determined by the selection equation:

$$y_{2i}^* = \beta_2' x_{2i} + u_{2i} \tag{2.20}$$

When the latent index $y_{2i}^*$ is positive, $y_{1i}$ is observed; otherwise $y_{1i}$ is censored. Once a distribution is chosen for the errors, the model defined by equations (2.19-20) is fully determined.

A concrete example may help to motivate the specification (2.19-20). Heckman (1974) analyzed female labor supply using this setup. The market wage level for a female worker ($y_{1i}$) depends upon various observable characteristics of the worker (education, age, experience, denoted by the vector $x_{1i}$) as well as various unobservable characteristics (represented by $u_{1i}$). However, many married women choose not to work outside the home, so any data on the wages of female workers is subject to considerable self-selection. Heckman modelled the labor force participation decision using a standard reservation wage model. Each woman sets a reservation wage level: if the woman finds an employer willing to offer a wage higher than the reservation wage, the woman accepts the wage offer and is employed. Let $y_{2i}^*$ denote the difference between the market wage offered to worker $i$ and her reservation wage. Presumably $y_{2i}^*$ would be affected by any variable affecting the market wage $y_{1i}$ as well as some factors irrelevant to the worker's productivity (marital status is one possible factor of this type), so $x_{2i}$ would include the elements of $x_{1i}$ as well as some additional

variables. When $y_{2i}^*$ is positive (or, equivalently, when $y_{2i} = 1$), then the market wage exceeds the reservation wage, the woman is employed, and her wage is observable. When $y_{2i}^*$ is negative, the woman is unemployed and $y_{1i}$ is censored.

Estimating (2.19) by applying least squares to the uncensored observations results in biased estimates for the same reasons that least squares fails in the Tobit model. That is, in the subsample of uncensored observations, the errors $u_{1i}$ have a non-zero mean, which can be shown to be:

$$E(u_{1i}|x_{1i}, x_{2i}, y_{2i}^* > 0) = \frac{\sigma_{12}}{\sigma_2}\lambda\left(\frac{\beta_2' x_{2i}}{\sigma_2}\right) \equiv \frac{\sigma_{12}}{\sigma_2}\lambda_i^*. \tag{2.21}$$

Equation (2.21) is a generalization of (2.15) that allows the censoring to be governed by a separate equation. It reduces to (2.15) when $y_{2i}^* = y_{1i}$.

To estimate the system of equations (2.19) and (2.20) by maximum likelihood methods requires a specification for the joint distribution of $(u_{1i}, u_{2i})$. It is conventional to assume that $(u_{1i}, u_{2i})$ are independent identically distributed with a bivariate normal distribution with mean zero, variances $\sigma_1^2$ and $\sigma_2^2$, and covariance $\sigma_{12}$. Since $y_{2i}^*$ in (2.20) is latent, we define the dummy variable $y_{2i} = 1$ if $y_{2i}^* > 0$ and $y_{2i} = 0$ otherwise. That is, $y_{1i}$ is observed if $y_{2i} = 1$ and otherwise is censored.

The model, as written, is not identified, since (2.20) can be multiplied by any positive number without affecting any of the observables. For example, divide (2.20) by $\sigma_2$:

$$\frac{y_{2i}^*}{\sigma_2} = \left(\frac{\beta_2}{\sigma_2}\right)' x_{2i} + \frac{u_{2i}}{\sigma_2}. \tag{2.22}$$

The sign of $y_{2i}^*/\sigma_2$ is the same as that of $y_{2i}^*$ so the implied value of $y_{2i}$ is unaffected. Insofar as the observable variables $y_{2i}$ and $x_{2i}$ are concerned, equations (2.20) and (2.21) are indistinguishable. Thus, the variance $\sigma_2^2$ is unidentified and can be set to any arbitrary value. A convenient normalization is $\sigma_2^2 = 1$. Then the probability that an observation is not censored (conditional on $x_{2i}$) is:

$$Q_i(\beta_2) = \Phi(\beta_2' x_{2i}), \tag{2.23}$$

while the component of the likelihood for an uncensored observation is

$$P_i(\beta_1, \beta_2, \sigma_1^2, \sigma_{12}) = \frac{1}{\sigma_1}\phi\left(\frac{y_{1i} - \beta_1' x_{1i}}{\sigma_1}\right)\Phi\left(\frac{\beta_2' x_{2i} + \sigma_{12}(y_{1i} - \beta_1' x_{1i})/\sigma_1^2}{\sqrt{1 - \sigma_{12}/\sigma_1^2}}\right) \tag{2.24}$$

It follows that the log-likelihood function is given by:

$$L(\beta_1, \beta_2, \sigma_1^2, \sigma_{12}) = \sum_{i=1}^{n} y_{2i} \log P_i(\beta_1, \beta_2, \sigma_1^2, \sigma_{12}) + (1 - y_{2i}) \log Q_i(\beta_2) \tag{2.25}$$

Thus, the likelihood function is relatively simple and only requires the numerical evaluation of one-dimensional normal integral $\Phi(\cdot)$ for which there are several good algorithms. Further discussion may be found in Griliches, Hall, and Hausman (1978).

Heckman (1979) proposed a simple two-step procedure for estimating the model (2.19-20) that avoids some of the complications of full ML estimation. In the first step of the Heckman procedure, $\beta_2$ is estimated by applying probit analysis to the selection equation alone. That is, one maximizes the marginal likelihood function for $y_{2i}$:

$$L_2(\beta_2) = \sum_{i=1}^{n} y_{2i} \log \Phi(\beta_2' x_{2i}) + (1 - y_{2i}) \log(1 - \Phi(\beta_2' x_{2i})) \tag{2.26}$$

Denote the first stage estimate of $\beta_2$ by $\hat{\beta}_2$. Heckman then suggests estimating $\lambda_i^*$ by:

$$\hat{\lambda}_i^* = \lambda(\hat{\beta}_2' x_{2i}) \tag{2.27}$$

For the uncensored observations, we have from (2.21) and the normalization $\sigma_2 = 1$:

$$y_{1i} = \beta_1' x_{1i} + \sigma_{12} \hat{\lambda}_i^* + (u_{2i} - \sigma_{12} \hat{\lambda}_i^*). \tag{2.28}$$

The error in (2.28) is heteroscedastic, but (asymptotically) uncorrelated with the right-hand side variables. Hence, applying least squares to (2.28) provides a consistent, though somewhat inefficient, estimator of $\beta_1$. Heckman (1979) explains how to obtain standard errors for the coefficients.

### Nonnormal Error Distributions

Our discussion has so far relied upon specification in which the errors are assumed to have a normal distribution. The econometric approach to selection bias is sometimes criticized for its dependence upon normality assumptions, but, in fact, normality is not an essential assumption. A variety of alternate parametric methods have been proposed to relax the normality assumption. Amemiya and Boskin (1974) considered the estimation of the Tobit model (2.2) when the errors have a log–normal rather than a normal distribution.

Dubin and McFadden (1984) consider estimation of the Heckman selectivity model (2.19–20) under the assumption that $u_{2i}$ has a logistic rather than a normal distribution. The selection equation then is of the logit rather than the probit form (see section 3 below). Further discussion of selection bias with parametric non-normal distributions is given in Lee (1982).

The estimation of the system of equations represented by (2.19) and (2.20) has generally relied on an assumed parametric form of the likelihood for the bivariate distribution of $(u_{1i}, u_{2i})$. However several researchers (Arabmazar and Schmidt (1981) and Goldberger (1983)), have pointed out that maximum likelihood estimation methods will yield inconsistent estimates of the parameters of interest if the parametric form of the error distribution is misspecified (whether it is assumed to be normal, log-normal, or logistic). Such misspecification may arise due to non–normality of the disturbance or may arise if maximum likelihood procedures are naively applied to aggregate data without consideration of heteroscedasticity. Since theory may not always suggest the proper parametric specification of the random disturbances, recent research in econometrics has focused on semiparametric methods.

Semiparametric methods seek identification and consistent estimation of the parameters of interest ($\beta_1$ in (2.19)) without a full-information specification of the selection equation. The bulk of the literature on semiparametric estimation of econometric models has considered the class of single disturbance models such as that presented in equation (2.2). Because these models involve only one error term, identification of the interest parameters can proceed under rather weak conditions, such as symmetry of the error distribution. (see Chamberlain, 1986). One simple semi-parametric estimator for the censored regression model is Powell's (1984) least absolute deviations (LAD) estimator. The logic of the LAD estimator is fairly simple. Consider equation (2.2) with only a constant term and no regressors. In this case the LAD estimator is median of the $y_i$'s (with censored observations replaced by zeroes). The least squares estimator is the mean of the $y_i$'s (again with censored observations replaced by zeroes). So long as less than half of the observations are censored, the median will be a consistent estimator of $\alpha$, while the sample mean will be downwardly biased. Powell shows that the same estimator is consistent when there are regressors.

Semiparametric estimation of the class of bivariate selection models given by (2.19) and (2.20) is not as well developed as that for the censored regression model (2.2). Heckman

and Robb (1985) have proposed a method of moments estimator, while Powell (1987) has extended recent work on semiparametric estimation of discrete choice models to this context. Estimation proceeds in two steps. First, an estimate of $\hat{\beta}_2$ is computed by applying semi-parametric methods to the selection equation alone. (Cosslett (1984), Stoker (1986), and others have suggested consistent estimators in this case.) In the second step, $\beta_1$ is estimated using semiparametric regression methods. The essential idea is that the conditional distribution of the errors $u_{1i}$ in equation (2.19) given the selection mechanism (2.20), depends only on $x_{2i}$ through the index $\beta_2' x_{2i}$. In the second step, the parameters of interest are identified through a comparison of pairs of observations for which the indices $\hat{\beta}_2' x_{2i}$ and $\hat{\beta}_2' x_{2j}$ are "close." See Powell (1987) for further discussion.

The development of semi-parametric methods is still at an early stage and we do not have much practical experience in the application of such methods. There is obviously a tradeoff between robustness and efficiency in the use of parametric and semi-parametric methods. We focus primarily on parametric methods that make fairly strong distributional assumptions, but it is mistaken to believe that the econometric approach to selectivity necessarily requires such assumptions.

## 3. Selection Bias in Binary Choice Models

Heckman's method provides a useful framework for handling linear regression models when the data are subject to an endogenous selection mechanism. Many applications in the social sciences, however, involve discrete dependent variables for which linear models are inappropriate. In this section, we discuss how the Heckman selection model can be adapted to models for dichotomous dependent variables. The most popular models of this sort are the logit and probit models. Before discussing selectivity corrections for these models, we briefly review the logit and probit specifications without the complications of censoring.

The most frequent occurence of dichotomous variables in the social sciences involves situations where a decision-maker faces a choice between two alternatives. The conventional model of choice in economics and other social sciences ascribes an unobservable level of utility $\tilde{U}_{ij}$ to alternative $j$ for decision-maker $i$. The primary purpose of most empirical studies of choice is to determine how various factors influence the attractiveness of the alternatives to different types of individuals. Although utility levels are unobservable (being analytical devices, rather than empirical measures), a regression-like framework provides

a convenient model for relating the attributes of the alternatives and decision-makers to utility levels:

$$\tilde{U}_{ij} = \beta_1' \tilde{x}_{ij} + \epsilon_{ij} \qquad (j = 1, 2) \tag{3.1}$$

where $\tilde{x}_{ij}$ usually includes the cost of alternative $j$ and other factors thought to affect choice. If utilities were observable, then regression methods could be applied directly to (3.1).

To estimate (3.1), it is necessary to invoke the hypothesis of utility maximization. A rational decision-maker should choose the alternative which maximizes his or her utility. Let $y_{1i}^*$ denote the difference between the utility of the first alternative and the second for decision-maker $i$:

$$y_{1i}^* = \tilde{U}_{i1} - \tilde{U}_{i2} = \beta_1' x_{1i} + u_{1i} \tag{3.2}$$

where $x_{1i} = \tilde{x}_{i1} - \tilde{x}_{i2}$ and $u_{1i} = \epsilon_{i1} - \epsilon_{i2}$. If $y_{1i}^* > 0$, the first alternative yields higher utility and is selected; otherwise the second alternative is selected. Define a dummy variable $y_{1i}$ denoting which alternative was selected:

$$y_{1i} = \begin{cases} 1, & \text{if } y_{1i}^* > 0; \\ 0, & \text{otherwise.} \end{cases} \tag{3.3}$$

At this point a convenient distribution is usually specified for the errors $\epsilon_{1i}$ and $\epsilon_{2i}$ and then the distribution of $y_{1i}$ is derived. From this point it is straightforward to obtain the maximum likelihood estimator for this model. (See Amemiya, 1984, for further discussion.)

The logit and probit models arise from different assumptions about the distribution of $\epsilon_{i1}$ and $\epsilon_{2i}$. If $\epsilon_{i1}$ and $\epsilon_{i2}$ are assumed to have independent type I extreme value distributions[5], then it can be shown that $u_{1i} = \epsilon_{i1} - \epsilon_{i2}$ has a *logistic distribution* with cdf:

$$F(u) = \frac{1}{1 + e^{-u}} \tag{3.4}$$

Alternatively, $\epsilon_{i1}$ and $\epsilon_{i2}$ can be assumed to have a joint normal distribution, each with mean zero, in which case the density of $u_{1i}$ is given by:

$$F(u) = \Phi(u/\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{u} e^{-t^2/2\sigma^2} dt \tag{3.5}$$

[5] The type I extreme value distribution has a cdf of the form $F(t) = \exp\{-e^{-t}\}$. See Johnson and Kotz (1970, chap. 21) for further discussion.

where $\sigma^2 = \text{Var}(\epsilon_{i1} + \epsilon_{i2})$. There is not much to choose between the two specifications. Both the logistic and normal distributions are symmetric and unimodal and, aside from different scale factors, differ only in their tails. Both have generalizations to choice models for more than two alternatives, but these will not concern us here.

There is no reason to believe selection bias is any less of a problem in logit and probit models than in linear regression models. However, its treatment is more difficult—at least computationally, if not conceptually—than in the linear model, so the possibility is frequently ignored. The remainder of this paper is devoted to the treatment of selection bias in binary choice models. From a conceptual point of view the development is entirely straightforward. One specifies a selection equation, resulting in a bivariate model that can, with appropriate distributional assumptions, be estimated by maximum likelihood. We examine the form of the likelihood equations and derive expressions for the information matrix. In sections 4 and 5 we specialize to the case of selection models of the probit and logit forms, respectively.

For binary choice models subject to selectivity, the specification is entirely analogous to the linear regression model of equations (2.19) and (2.20), except that the observable dependent variable in the outcome equation (2.19) is replaced by the latent variable formulation of equation (3.2). Following equation (2.20), we again specify a selection equation of the form:

$$y_{2i}^* = \beta_2' x_{2i} + u_{2i} \tag{3.6}$$

so that $y_{1i}$ is observed if and only if $y_{2i}^* > 0$. The corresponding indicator of whether $y_{1i}$ is observed or censored is again denoted $y_{2i}$:

$$y_{2i} = \begin{cases} 1, & \text{if } y_{2i}^* > 0; \\ 0, & \text{if } y_{2i}^* \leq 0. \end{cases} \tag{3.7}$$

The specification of (3.2) and (3.6) is the natural way to adapt Heckman's selection model to a binary choice situation.

In the case of the linear regression model with censoring described in section 2, a bivariate normal distribution is usually for the errors $u_{1i}$ and $u_{2i}$. The same assumption applied to the binary choice model (3.2) with selection equation (3.6) leads to a probit model with censoring. This case is covered in some detail in section 4. Alternatively, if we assume that $u_{1i}$ and $u_{2i}$ have a bivariate logistic distribution, then we obtain the logit model with censoring of section 5. The logit case is less clearcut than the probit, because

there are several possible choices for a bivariate logistic model. The parameterization we propose is flexible and computationally tractible.

Before specializing to particular distributions, we consider the ML estimator for case of an arbitrary bivariate distribution of $u_{1i}$ and $u_{2i}$. The following assumptions will be made:

**A1.** $(x_{1i}, x_{2i})$ is independent of $(u_{1i}, u_{2i})$. The cumulative distribution function of $(u_{1i}, u_{2i})$ is $F(u_{1i}, u_{2i})$.

**A2.** The observations $(x_{1i}, x_{2i}, u_{1i}, u_{2i})$ are independently and identically distributed.

Assumption A1 is that the explanatory variables be exogenously determined. Lee (1981) discusses estimation of selection models with endogenous regressors. Assumption A2 is that, aside from the censoring of some observations according to the selection rule (3.7), the observations were obtained by random sampling from some population. As most selectivity models are applied to cross-sectional survey data, this assumption should be satisfied at least approximately.

The choice of $F$, as emphasized above, is more a matter of computational convenience than anything else. $F$ should be sufficiently flexible to capture plausible forms of dependence between $u_{1i}$ and $u_{2i}$, but, if this requirement is satisfied, a simple parameterization should be the main concern. We will impose two restrictions on $F$. First, note that the location and scale parameters for $u_{1i}$ and $u_{2i}$ can be normalized to convenient values by appropriate shifts and rescalings of $y_{1i}^*$ and $y_{2i}^*$ as in the usual binary choice situation. Thus, there is little loss of generality in requiring that $F$ have identical marginal distributions for $u_1$ and $u_2$. Second, we will restrict ourselves to one parameter families for $F$ and will denote the parameter by $\rho$. In the normal case, $\rho$ will be the correlation between $u_{1i}$ and $u_{2i}$, while in the logit case the relationship is somewhat more complicated. To summarize, the joint cdf takes the form $F(u_1, u_2; \rho)$ and has marginal distbutions $H(u_1) \equiv F(u_1, \infty; \rho)$ and $F(\infty, u_2; \rho) \equiv H(u_2)$ which do not depend on $\rho$.[6]

Next, we calculate the probability of the three possible outcomes: a censored observation ($y_{2i} = 0$), an uncensored success ($y_{1i} = 1$ and $y_{2i} = 1$), and an uncensored failure

[6] Mardia (1970) discusses a general method for forming bivariate distributions with specified marginal distributions.

($y_{1i} = 0$ and $y_{2i} = 1$). This requires some additional notation. Let $G(\cdot,\cdot;\rho)$ denote the upper tail probability of $F(\cdot,\cdot;\rho)$, i.e.:

$$\begin{aligned} G(u_1,u_2;\rho) &= \Pr(u_{1i} > u_1, u_{2i} > u_2) \\ &= 1 - H(u_1) - H(u_2) + F(u_1,u_2;\rho) \end{aligned} \tag{3.8}$$

Then the probability of an observation *not* being censored is given by:

$$Q_i(\beta_2) = \Pr(y_{2i} = 1|x_{1i},x_{2i}) = \Pr(y_{2i}^* > 0) = 1 - H(-\beta_2' x_{2i}) \tag{3.9}$$

The probability of an uncensored success is given by:

$$\begin{aligned} P_i(\beta_1,\beta_2,\rho) &= \Pr(y_{1i} = 1, y_{2i} = 1|x_{1i},x_{2i}) \\ &= \Pr(y_{1i}^* > 0, y_{2i}^* > 0|x_{1i},x_{2i}) \\ &= G(-\beta_1' x_{1i}, -\beta_2' x_{2i}) \end{aligned} \tag{3.10}$$

Finally, the probability of an uncensored failure is given by:

$$Q_i(\beta_2) - P_i(\beta_1,\beta_2,\rho). \tag{3.11}$$

Combining (3.9–11) we obtain the log likelihood function:

$$\begin{aligned} L(\beta_1,\beta_2,\rho) = \sum_{i=1}^{n} y_{2i}\big(&y_{1i}\log P_i(\beta_1,\beta_2,\rho) \\ &+ (1-y_{1i})\log(Q_i(\beta_2) - P_i(\beta_1,\beta_2,\rho)) + (1-y_{2i})\log(1-Q_i(\beta_2)\big) \end{aligned} \tag{3.12}$$

The ML estimator of $\theta = (\beta_1,\beta_2,\rho)$ is obtained by maximizing (3.12) with respect to $\theta$. This is a somewhat more difficult maximization problem than the usual binary choice problem because $P_i(\beta_1,\beta_2,\rho)$ requires the computation of a bivariate integral. It is possible, however, to obtain some simplification of the optimization problem as shown below.

The first order conditions for the ML estimator are:

$$\frac{\partial L}{\partial \beta_1} = \sum_{i=1}^{n} y_{2i} \frac{y_{1i}Q_i(\beta_2) - P_i(\beta_1,\beta_2,\rho)}{P_i(\beta_1,\beta_2,\rho)\,(Q_i(\beta_2) - P_i(\beta_1,\beta_2,\rho))} \frac{\partial P_i(\beta_1,\beta_2,\rho)}{\partial \beta_1} \tag{3.13}$$

$$\frac{\partial L}{\partial \beta_2} = \sum_{i=1}^{n} \frac{y_{2i} - Q_i(\beta_2)}{Q_i(\beta_2)\,(1-Q_i(\beta_2))} \frac{\partial Q_i(\beta_2)}{\partial \beta_2}$$

$$- \sum_{i=1}^{n} \frac{y_{2i}}{Q_i(\beta_2) - P_i(\beta_1, \beta_2, \rho)} \Bigg[ \left( 1 - \frac{y_{1i} Q_i(\beta_2)}{P_i(\beta_1, \beta_2, \rho)} \right) \frac{\partial P_i(\beta_1, \beta_2, \rho)}{\partial \beta_2}$$

$$+ \left( y_{1i} - \frac{P_i(\beta_1, \beta_2, \rho)}{Q_i(\beta_2)} \right) \frac{\partial Q_i(\beta_2)}{\partial \beta_2} \Bigg] \tag{3.14}$$

$$\frac{\partial L}{\partial \rho} = \sum_{i=1}^{n} y_{2i} \frac{y_{1i} Q_i(\beta_2) - P_i(\beta_1, \beta_2, \rho)}{P_i(\beta_1, \beta_2, \rho) \left( Q_i(\beta_2) - P_i(\beta_1, \beta_2, \rho) \right)} \frac{\partial P_i(\beta_1, \beta_2, \rho)}{\partial \rho} \tag{3.15}$$

Some insight into the first order conditions (3.13–15) can be obtained by noting that:

$$\frac{y_{1i} Q_i(\beta_2) - P_i(\beta_1, \beta_2, \rho)}{P_i(\beta_1, \beta_2, \rho)(Q_i(\beta_2) - P_i(\beta_1, \beta_2, \rho))} = \frac{y_{1i} - R_i(\beta_1, \beta_2, \rho)}{R_i(\beta_1, \beta_2, \rho)(1 - R_i(\beta_1, \beta_2, \rho))}, \tag{3.16}$$

where $R_i(\beta_1, \beta_2, \rho) = P_i(\beta_1, \beta_2, \rho)/Q_i(\beta_2)$ is the conditional probability $y_{1i} = 1$ given $y_{2i} = 1$. Thus, the first order conditions essentially "fit" the uncensored observations on $y_{1i}$ to their conditional expectation $R_i(\beta_1, \beta_2, \rho)$.

Equations (3.13–15) are a system of nonlinear equations that can be difficult to solve numerically, though the computational requirements are not impossible. An alternative two-step estimation procedure is available which allows some simplification in computation at the cost of a reduction of the efficiency of the resulting estimators. The first line of equation (3.14) is the first order condition from a binary choice model without censoring. The term inside the square brackets on the second and third lines of (3.14) has expectation zero conditional on $x_{1i}$, $x_{2i}$, and $y_{2i} = 1$. Thus, if we neglect this term, we can obtain a consistent estimator of $\beta_2$ by solving:

$$\sum_{i=1}^{n} \frac{y_{2i} - Q_i(\hat{\beta}_2)}{Q_i(\hat{\beta}_2) \left( 1 - Q_i(\hat{\beta}_2) \right)} \frac{\partial Q_i(\hat{\beta}_2)}{\partial \beta_2} = 0 \tag{3.17}$$

for $\hat{\beta}_2$. This amounts to either a logit or probit analysis of the selection equation *alone*. In the second step of the estimation procedure, one then solves equations (3.13) and (3.15) for $\beta_1$ and $\rho$ after replacing $\beta_2$ by $\hat{\beta}_2$. Notice that equations (3.13) and (3.15) only involve the uncensored observations and have a structure similar to that of the usual binary choice problem. The standard errors obtained in the second step must be corrected for the estimation of $\hat{\beta}_2$ in the first step. (See Vuong, 1985, or Duncan, 1987, for details.)

## 4. Estimation in the Normal Case

The results in the previous section are easily specialized to the case where the errors have a standardized bivariate normal distribution with correlation coefficient $\rho$. In the censored probit model, the joint cdf of $u_{1i}, u_{2i}$ is assumed to be:

$$F(u_{1i}, u_{2i}; \rho) = \frac{1}{2\pi} \int_{-\infty}^{u_{1i}} du_1 \int_{-\infty}^{u_{2i}} \exp\{-\frac{1}{2(1-\rho^2)}(u_1^2 - \rho u_1 u_2 + u_2^2)\} du_2 \tag{4.1}$$

The probability of an observation being uncensored (conditional on $x_{1i}$ and $x_{2i}$) is given by:

$$Q_i(\beta_2) = \Phi(\beta_2' x_{2i}) \tag{4.2}$$

and the probability of an uncensored success is given by:

$$P_i(\beta_1, \beta_2, \rho) = F(\beta_1' x_{1i}, \beta_2' x_{2i}, \rho) \tag{4.3}$$

where we have used the fact that $\Phi(x) = 1 - \Phi(-x)$ and $G(x, y, \rho) = F(-x, -y, \rho)$. The gradients in the probit case also take a fairly simple form:

$$\frac{\partial P_i(\beta_1, \beta_2, \rho)}{\partial \beta_1} = \phi(\beta_1' x_{1i}) \Phi\left(\frac{\beta_2' x_{2i} - \rho \beta_1' x_{1i}}{\sqrt{1-\rho^2}}\right) x_{1i} \tag{4.4}$$

$$\frac{\partial P_i(\beta_1, \beta_2, \rho)}{\partial \beta_2} = \phi(\beta_2' x_{2i}) \Phi\left(\frac{\beta_1' x_{1i} - \rho \beta_2' x_{2i}}{\sqrt{1-\rho^2}}\right) x_{2i} \tag{4.5}$$

$$\frac{\partial Q_i(\beta_2)}{\partial \beta_2} = \phi(\beta_2' x_{2i}) x_{2i} \tag{4.6}$$

$$\frac{\partial P_i(\beta_1, \beta_2, \rho)}{\partial \rho} = (1-\rho^2)^{-1/2} \phi(\beta_1' x_{1i}) \phi\left(\frac{\beta_2' x_{2i} - \rho \beta_1' x_{1i}}{\sqrt{1-\rho^2}}\right) \tag{4.7}$$

Subsituting (4.4-7) into (3.13-15) and solving provides ML estimates for the probit model with selectivity.

Computation of the ML estimates is a non-trivial problem. It would be convenient to have a way of testing for the presence of selection bias without having to compute the ML estimates. The null hypothesis of no selection bias is $H_0 : \rho = 0$. There are a variety of ways of testing this hypothesis which have the same asymptotic properties. Wald's method,

for instance, would require that we estimate $\rho$ by maximum likelihood and compute the statistic:

$$W = \hat{\rho}^2 / \hat{V}(\hat{\rho}), \tag{4.8}$$

which has an approximate chi-square distribution with one degree of freedom under the null-hypothesis. (The statistic in (4.8) is the square of the usual $t$-statistic for testing $\rho = 0$.) The likelihood ratio statistic compares the value of the likelihood function at the ML and constrained ML estimates. The constrained ML estimates are easily obtained since, when $\rho = 0$:

$$P_i(\beta_1, \beta_2, 0) = \Phi(\beta_1' x_{1i}) Q_i(\beta_2' x_{2i}). \tag{4.9}$$

In this case, the constrained ML estimates can be obtained from two univariate probit analyses. First, estimate the outcome equation (using the non-missing observations) to obtain the constrained ML estimate $\tilde{\beta}_1$. Second, estimate the selection equation by probit analysis to obtain the constrained ML estimates $\tilde{\beta}_2$. The value of the log likelihood evaluated at the constrained ML estimates, denoted $L(\tilde{\beta}_1, \tilde{\beta}_2, 0)$, is the sum of the log likelihoods from the two univariate probit analyses. The LR statistic for testing $\rho = 0$ is:

$$LR = -2\left(L(\tilde{\beta}_1, \tilde{\beta}_2, 0) - L(\hat{\beta}_1, \hat{\beta}_2, \hat{\rho})\right), \tag{4.10}$$

which also has an approximate chi-square distribution with one degree of freedom.

The disadvantage of the Wald and likelihood ratio statistics is that both require computation of the full ML estimates. We have shown, however, that the constrained ML estimates are easily obtained and do not require specialized software. An easier method that avoids computation of the full model is the score test procedure (Rao, 1973: 417–18). If $\rho = 0$, the gradients (3.13–15) simplify to:

$$\frac{\partial L}{\partial \beta_1} = \sum_{i=1}^{n} y_{2i} x_{1i} \hat{u}_{1i} \tag{4.11}$$

$$\frac{\partial L}{\partial \beta_2} = \sum_{i=1}^{n} x_{2i} \hat{u}_{2i} \tag{4.12}$$

$$\frac{\partial L}{\partial \rho} = \sum_{i=1}^{n} \hat{u}_{1i} \hat{u}_{2i} \tag{4.13}$$

where $\hat{u}_{1i}$ and $\hat{u}_{2i}$ are "generalized residuals" (Gourieroux et al., 1987):

$$\hat{u}_{1i} = \frac{\phi(\tilde{\beta}_1' x_{1i})}{\Phi(\tilde{\beta}_1' x_{1i})(1 - \Phi(\tilde{\beta}_1' x_{1i}))}(y_{1i} - \Phi(\tilde{\beta}_1' x_{1i})) \tag{4.14}$$

$$\hat{u}_{2i} = \frac{\phi(\tilde{\beta}_2' x_{2i})}{\Phi(\tilde{\beta}_2' x_{2i})(1 - \Phi(\tilde{\beta}_2' x_{2i}))}(y_{2i} - \Phi(\tilde{\beta}_2' x_{2i})) \tag{4.15}$$

If the null hypothesis is correct, the gradients (4.11–13) should be close to zero. The score statistic is a quadratic form in the gradients with the information matrix (or a consistent estimate) as weighting matrix. A convenient method for obtaining the score statistic is to perform an "artificial regression" where the dependent variable equals one for all observations and the independent variables are $y_{2i}x_{1i}\hat{u}_{1i}$, $x_{2i}\hat{u}_{2i}$, and $\hat{u}_{1i}\hat{u}_{2i}$. Computed in this way, the score statistic is:

$$S = nR^2 \tag{4.16}$$

where the $R^2$ is obtained from the artificial regression described above.

## 5. Estimation in the Logistic Case

The main task in specializing to the logistic case is to choose a bivariate logistic distribution. The usual suggestions for a bivariate logistic distribution allow only very restricted forms of correlation (see Johnson and Kotz, 1972, pp. 291-94). We propose an alternative bivariate logistic distribution that is an improvement by this criterion:

$$F(u_1, u_2, \rho) = \frac{1}{1 + \left(e^{-u_1/\rho} + e^{-u_2/\rho}\right)^{1/\rho}} \tag{5.1}$$

where the parameter $\rho$ can only take positive values. $F$ is in fact a bivariate logistic distribution as its marginals are of the logistic form, e.g.:

$$H(u_1) = \lim_{u_2 \to \infty} F(u_1, u_2) = \frac{1}{1 + e^{-u_1}} \tag{5.2}$$

It can be shown that for $0 < \rho \leq \sqrt{2}$ that $\text{corr}(u_1, u_2) = 1 - \rho^2/2$ so the case $\rho = \sqrt{2}$ corresponds to no correlation between $u_1$ and $u_2$. A zero correlation between $u_1$ and $u_2$, however, does not imply that they are independent and, in fact, for no value of $\rho$ will $u_1$ and $u_2$ be independent. With these reservations noted, we proceed to develop a logit model with selection.

From equations (3.8-10), substituting (5.1-2), we obtain the necessary probabilities to form the likelihood:

$$Q_i(\beta_2) = \frac{1}{1+e^{-\beta_2' x_{2i}}} \tag{5.3}$$

$$P_i(\beta_2) = 1 - \frac{1}{1+e^{-\beta_1' x_{1i}}} - \frac{1}{1+e^{-\beta_2' x_{2i}}} + \frac{1}{1+\left(e^{\beta_1' x_{1i}/\rho} + e^{\beta_2' x_{2i}/\rho}\right)^{1/\rho}} \tag{5.4}$$

The probabilities in (5.3-4) are somewhat easier to compute than in the normal case, but the expressions for the derivatives are more complex:

$$\frac{\partial P_i(\beta_1,\beta_2,\rho)}{\partial \beta_1} = \Big(H(\beta_1' x_{1i})(1-H(\beta_1' x_{1i})) - \rho^{-2}\frac{e^{\beta_1' x_{1i}/\rho}}{e^{\beta_1' x_{1i}} + e^{\beta_2' x_{2i}}} \times$$

$$F(-\beta_1' x_{1i}, -\beta_2' x_{2i})(1 - F(-\beta_1' x_{1i}, -\beta_2' x_{2i}))\Big) x_{1i} \tag{5.5}$$

$$\frac{\partial P_i(\beta_1,\beta_2,\rho)}{\partial \beta_2} = \Big(H(\beta_2' x_{2i})(1-H(\beta_2' x_{2i})) - \rho^{-2}\frac{e^{\beta_2' x_{2i}/\rho}}{e^{\beta_1' x_{1i}} + e^{\beta_2' x_{2i}}} \times$$

$$F(-\beta_1' x_{1i}, -\beta_2' x_{2i})(1 - F(-\beta_1' x_{1i}, -\beta_2' x_{2i}))\Big) x_{2i} \tag{5.6}$$

$$\frac{\partial P_i(\beta_1,\beta_2,\rho)}{\partial \rho} = \frac{1}{\rho^2}\left(1 + \log(e^{\beta_1' x_{1i}/\rho} + e^{\beta_2' x_{2i}/\rho})\right) \times$$

$$F(\beta_1' x_{1i}, \beta_2' x_{2i}, \rho)\,(1 - F(\beta_1' x_{1i}, \beta_2' x_{2i}, \rho)) \tag{5.7}$$

$$\frac{\partial Q_i(\beta_2)}{\partial \beta_2} = H(\beta_2' x_{2i})(1 - H(\beta_2' x_{2i}))x_{2i} \tag{5.8}$$

Once more, subsituting (5.5–8) into (3.13–15) and solving provides ML estimates for the logit model with selectivity.

## 6. Empirical Application: Turnout and Voting Behavior

As an application of the methods described in the preceding sections, we consider the analysis of political preferences using voting data. Voting behavior has been of interest not only to political scientists, but also to sociologists, psychologists, and economists because voting reflects a variety of social, psychological, and economic concerns. As diverse as these approaches are, they share a common structure: the characteristics of voters determine their group memberships, attitudes, or preferences, and vote choices are taken to be a

measure of such memberships, attitudes, or preferences. Our purpose here is not to engage in a debate over which approach to voting analysis is superior, but only to point out that virtually *all* such analyses are subject to selection problems.

Empirical voting research is primarily concerned with the relation between various political, demographic, and psychological characteristics and political preferences. In two-candidate elections, vote and candidate preference are synonymous (unlike multicandidate elections, where strategic factors may make it in a voter's interest to vote for someone other than his or her most preferred candidate), so there appears to be little point in distinguishing between vote and preference. Nonvoters, however, also have preferences, but they do not vote. If preference is measured by vote, then data on preference is missing for non-voters.

In the U.S. voting literature, vote equations are invariably interpreted in terms of preferences and attitudes. Presumably the same model of preference applies to non-voters as well as voters. If turnout and preference are unrelated, there should be no bias in estimating a model of preference based on the subsample of voters whose preference is observed. To the degree that there are common factors determing both turnout and preference, turnout is a source of selection bias.

The customary practice in voting studies has been to analyze turnout and vote choice separately. The voting electorate, however, is not a random subsample of the voting age population. Voters are known to be older, more educated, and more likely to be married than non-voters (Wolfinger and Rosenstone, 1980). The effects of race and gender are less clear. Blacks vote at lower rates than whites, but it has been argued that black turnout is as high or higher than white turnout after controlling for education and income (Wolfinger and Rosenstone, 1980: 90–91; Verba and Nie, 1972: 170–71). In the 1950's male turnout was approximately ten points higher than female turnout (Campbell et al., 1960: 485–89), but the gender gap in turnout has eroded considerably since then (Wolfinger and Rosenstone, 1980: 41–44) to the point that women may now participate at slightly higher levels than men. Registration laws tend to reduce turnout rates among the more mobile segments of the population (Squire, Wolfinger, and Glass, 1986). On the other hand, Wolfinger and Rosenstone (1980: 109–13) argue that there are no significant ideological differences between voters and non-voters.

Some of the variables that appear in turnout studies are clearly relevant to preference. Blacks vote overwhelmingly Democratic. The gender gap in Republican support has been

widely discussed, as have generational differences. Other variables that influence voting rates, such as education or residential mobility, do not correspond very closely to any current cleavage in American politics and can be safely omitted from a vote equation.

Data from the 1984 American National Election Study (NES) were used to estimate probit models of vote and turnout. The outcome variable is whether the respondent voted for Ronald Reagan and is missing for non-voters. In this case, the selection equation is a standard turnout equation. In the NES survey there is some overreporting of turnout. After the post-election interview, public voting records were examined to determine whether respondent's who claimed to have voted actually did and our analysis is based on the "validated" turnout variable. Estimating a vote equation using only validated voters should produce results similar to those based on an exit poll.

For purposes of comparison, we present in Table 1 separate probit analyses of vote and turnout. The vote equation in the first column of Table 1 includes 1347 validated voters. Blacks, women, union households, persons over 55 years old, and self-classified liberals were more less likely to vote for Reagan, though gender was insignificant and age only marginally significant. Estimates for the turnout equation are presented in the second column of Table 1. Respondents who have lived at their current address for less than a year were classified as "new residents" and were found to turn out at much lower rates, as were younger voters and blacks. Respondents who had attended college, were married, either read a newspaper or watched network evening news on a daily basis, or belonged to a labor household were more likely to turnout. Women were slightly more likely to vote than men.

Are the estimates of the vote equation in Table 1 subject to selection bias? The score test described in section 4 was performed and the null hypothesis could be rejected ($X^2 = 4.32$ with one degree of freedom, $p < 0.01$). The likelihood ratio and Wald statistics were 3.78 and 5.24, respectively.

The model in Table 1 was reestimated using the bivariate normal selection model of section 4. Maximum likelihood estimates are presented in Table 2. The estimated turnout equation in the second column of Table 2 is, for all practical purposes, identical to that in Table 1, as should be the case if the bivariate model is correctly specified. The estimated coefficients in the outcome equation, however, do change after the correction for self-selection. The largest differences between Tables 1 and 2 are in the age coefficients. After correcting for turnout, we find a much stronger relationship between age and Reagan

preference (with younger voters more likely to prefer Reagan) and the estimated gender gap is larger and significant (for a one-tailed test with a 0.05 significance level). The coefficients of the ideology dummies are slightly smaller than those reported in Table 1.

The estimated correlation between the errors in the turnout and vote equations is -0.41 which implies that, after controlling for measured characteristics, non-voters were more likely to prefer Reagan than voters. Our estimates suggests that the Democratic loss in 1984 is not attributable to low turnout. Note also that the estimated intercept increases substantially after correcting for self-selection.

## 7. Conclusion

Missing data problems are pervasive in the social sciences. The econometric approach to selectivity, pioneered by Heckman (1979), provides a useful framework for modelling self-selection mechanisms. The econometric approach relies on an economic or other social scientific theory for guidance in modelling the selection process, but if one is willing to subscribe to some specification—as we suspect most social scientists are willing to do—it allows most missing data problems to be overcome. The main contribution of this article was to indicate how the Heckman model could be extended to probit and logit models. The test for selection bias in the probit model (described in section 4) is suggested as a useful diagnostic for situations when the selection problem is not the primary focus of attention.

## Appendix A

### Derivation of Equation (2.10)

It is fairly straightforward to prove that the sign of $\pi$ is the opposite of that of $\beta$. First, a bit of notation. Let $F$ denote the cumulative distribution function (c.d.f.) of $u_i$ and assume that $F$ is continuously differentiable with density $f = F'$. Let $g$ denote the density of $x_i$ and assume $x_i$ and $u_i$ are independent. To avoid unnecessary technical details, suppose that $f(u) > 0$ for all $u$. Define:

$$\xi(t) = E(u_i|u_i > t) = \frac{1}{1 - F(t)} \int_t^\infty u f(u) du. \tag{A.1}$$

Note that:

$$\begin{aligned} \xi'(t) &= \frac{f(t)}{(1 - F(t))^2} \int_t^\infty u f(u) du - \frac{t f(t)}{1 - F(t)} \\ &= h(t)(\xi(t) - t) \end{aligned} \tag{A.2}$$

where $h(t) = f(t)/(1 - F(t))$ is the *hazard function.* (An interpretation of the hazard function is that $h(x)dx$ is the conditional probability of a random variable $X$ with density $f(x)$ falling in the interval $(x, x + dx)$ given that $X > x$.) It follows from (A.2) that $\xi'(t) \geq 0$ for all $t$.

Since $x$ and $u$ are assumed to be independent (in the full sample), it follows that:

$$\xi_i = \xi(c - \alpha - \beta x_i). \tag{A.3}$$

Let $\mu = E(x_i|y_i \geq c)$. Then, expanding $\xi(t)$ around the point $t = c - \alpha - \beta\mu$, by the mean value theorem there exists $z(x)$ between $x$ and $\mu$ such that:

$$\xi_i = \xi(c - \alpha - \beta\mu) - \beta(x - \mu)\xi'(c - \alpha - \beta z(x)) \tag{A.4}$$

Hence:

$$\begin{aligned} \text{Cov}(x_i, \xi_i|y_i > c) &= \int (x - \mu)\xi(c - \alpha - \beta x) g(x|y \geq c) dx \\ &= \xi(c - \alpha - \beta\mu) \int (x - \mu) g(x|y \geq c) dx \\ &\quad - \beta \int (x - \mu)^2 \xi'(c - \alpha - \beta z(x)) g(x|y \geq c) dx \\ &= -\beta \int (x - \mu)^2 \xi'(c - \alpha - \beta z(x)) g(x|y \geq c) dx \end{aligned} \tag{A.5}$$

using (A.2). The integrand on the last line of (A.5) is non-negative, so the sign of $\mathrm{Cov}(x_i, \xi_i))$ is the opposite that of $\beta$ except, of course, when $\beta = 0$ and the covariance is zero.

## Table 1

Probit Estimates of Vote and Turnout

| Variable | Outcome Equation (Reagan Vote) | Selection Equation (Turnout) | Sample Mean (Voters Only) |
|---|---|---|---|
| Constant | 0.18 | -0.29 | |
| | (0.09) | (0.09) | |
| Black | -1.37 | -0.27 | 0.08 |
| | (0.18) | (0.09) | |
| Female | -0.09 | 0.14 | 0.57 |
| | (0.07) | (0.06) | |
| Union | -0.51 | 0.20 | 0.24 |
| | (0.09) | (0.07) | |
| Under 30 | 0.03 | -0.22 | 0.21 |
| | (0.10) | (0.07) | |
| Over 55 | -0.19 | 0.18 | 0.33 |
| | (0.09) | (0.07) | |
| Liberal | -0.40 | — | -0.19 |
| | (0.10) | | |
| Conservative | 0.52 | — | 0.32 |
| | (0.08) | | |
| New Resident | — | -0.53 | 0.14 |
| | | (0.07) | |
| College | — | 0.62 | 0.49 |
| | | (0.07) | |
| Married | — | 0.26 | 0.62 |
| | | (0.06) | |
| TV/Newspaper Usage | — | 0.32 | 0.70 |
| | | (0.06) | |
| Log Likelihood | -817 | -1344 | |
| $n$ | 1347 | 2237 | |

## Table 2

Bivariate Normal Selection Model

| Variable | Outcome Equation (Reagan Vote) | Selection Equation (Turnout) | Sample Mean (Full Sample) |
|---|---|---|---|
| Constant | 0.49 | -0.29 | |
| | (0.11) | (0.09) | |
| Black | -1.22 | -0.27 | 0.11 |
| | (0.20) | (0.09) | |
| Female | -0.11 | 0.14 | 0.56 |
| | (0.07) | (0.06) | |
| Union | -0.55 | 0.20 | 0.21 |
| | (0.08) | (0.07) | |
| Under 30 | 0.14 | -0.22 | 0.28 |
| | (0.10) | (0.07) | |
| Over 55 | -0.24 | 0.19 | 0.29 |
| | (0.08) | (0.07) | |
| Liberal | -0.36 | — | -0.18 |
| | (0.09) | | |
| Conservative | 0.49 | — | 0.29 |
| | (0.08) | | |
| New Resident | — | -0.53 | 0.21 |
| | | (0.07) | |
| College | — | 0.62 | 0.41 |
| | | (0.06) | |
| Married | — | 0.28 | 0.57 |
| | | (0.06) | |
| TV/Newspaper Usage | — | 0.31 | 0.62 |
| | | (0.06) | |
| $\rho$ | | -0.41 | |
| | | (0.14) | |
| Log Likelihood | | -2159 | |
| $n$ | | 2237 | |

## Biographical Sketches

**Jeffrey A. Dubin** is Associate Professor of Economics at the California Institute of Technology. He received his Ph.D. from the Massachusetts Institute of Technology in 1982. He is the author of *Consumer Durable Choice and the Demand for Electricity* (North-Holland, 1985) and articles in the *American Economic Review*, *Econometrica*, and other journals.

**Douglas Rivers** is Associate Professor of Political Science at the University of California, Los Angeles. He received his Ph.D. from Harvard University in 1981. He has contributed articles to the *American Political Science Review*, *American Journal of Political Science*, *Journal of Econometrics*, and other journals.

## References


AMEMIYA, T. (1984), "Tobit Models: A Survey," *Journal of Econometrics*, 24 No. 1.

AMEMIYA, T., and M. BOSKIN (1974), "Regression Analysis When the Dependent Variable Is Truncated Lognormal, with an Application to the Determinants of the Duration of Welfare Dependency," *International Economic Review* 15: 485–496.

ARABMAZAR, A. and P. SCHMIDT (1981), "Further Evidence on the Robustness of the Tobit Estimator to Heteroscedasticity," *Journal of Econometrics*, 17: 253–258.

BILLINGSLEY, P. (1987), *Probability and Measure*, 2nd ed., New York: John Wiley.

CAMPBELL, A., P.E. CONVERSE, W. E. MILLER, and D. E. STOKES (1960), *The American Voter*, New York: John Wiley & Sons.

CHAMBERLAIN, G. (1986a), "Asymptotic Efficiency in Semiparametric Models with Censoring," *Journal of Econometrics*, 32: 189–218.

COSSLETT, S.R. (1984), "Distribution–Free Maximum Likelihood Estimator of the Binary Choice Model," *Econometrica* 51: 765–782.

DUBIN, J.A., and D.L. McFADDEN (1984), "Econometric Analysis of Residential Electric Appliance Holdings and Consumption," *Econometrica*, 52 No. 2.

DUBIN, J.A., and R.D. RIVERS (1989), *Statistical Software Tools*, 2nd ed. Pasadena: Dubin/Rivers Research.

DUNCAN, G. M. (1987) "A Simplified Approach to $M$-estimation with Application to Two-Stage Estimators," *Journal of Econometrics*, 34: 373–90.

GOLDBERGER, A.S. (1983), "Abnormal Selection Bias," in S. Karlin, *et. al.*, eds., *Studies in Econometrics, Time Series, and Multivariate Statistics.* New York: Academic Press.

GRILICHES, Z. (1957), "Specification Bias in Estimates of Production Functions," *Journal of Farm Economics*, 39: 8-20.

GRILICHES, Z., B.H. HALL, and J.A. HAUSMAN (1978), "Missing Data and Self–Selection in Large Panels," *Annales De L'Insee* - N ° 30–31.

HECKMAN, J.J. (1974), "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, 42.

HECKMAN, J.J. (1979), "Sample Selection Bias as a Specification Error," *Econometrica*, 47 No. 1.

HECKMAN, J.J., and R. ROBB (1985) "Alternative Methods for Evaluating the Impact of Interventions," in J.J. Heckman and B. Singer, eds., *Longitudinal Analysis of Labor Market Data.* New York: Cambridge University Press.

JOHNSON, N.L., and S. KOTZ (1970), *Distributions in Statistics: Continuous Univariate Distributions—I*, New York: John Wiley.

JOHNSON, N.L., and S. KOTZ (1972), *Distributions in Statistics: Continuous Multivariate Distributions*, New York: John Wiley.

LEE, L.F. (1981), "Simultaneous Equations Models with Discrete Endogenous and Censored Variables," in C. F. Manski and D. McFadden, eds., *Structural Analysis of Discrete Data with Econometric Applications.* Cambridge: MIT Press.

LEE, L.F. (1982), "Some Approaches to the Correction of Selectivity Bias," *Review of Economic Studies* 49: 355–372.

LITTLE, R.J.A., and D.B. RUBIN. (1987), *Statistical Analysis with Missing Data,* New York: John Wiley.

MARDIA, K.V. (1970) *Families of Bivariate Distributions*, London: Hafner.

POWELL, J.L. (1984), "Least Absolute Deviations Estimation for the Censored Regression Model," *Journal of Econometrics*, 25: 303–325.

POWELL, J.L. (1987), "Semiparametric Estimation of Bivariate Latent Variable Models," Social Systems Research Institute, University of Wisconsin Working Paper No. 8704.

RAO, C.R. (1973) *Linear Statistical Inference and Its Applications*, 2nd ed. New York: John Wiley.

SQUIRE, P., R.E. WOLFINGER, and D.P. GLASS (1987), "Residential Mobility and Voter Turnout," *American Political Science Review*, 81: 45-65.

STOKER, T.M. (1986), "Consistent Estimation of Scaled Coefficients," *Econometrica*, 54: 1461–1481.

THEIL, H. (1957), "Specification Errors and the Estimation of Economic Relationships," *Review of the International Statistical Institute*, 25: 41-51.

TOBIN, J. (1958), "Estimation of Relationships for Limited Dependent Variables," *Econometrica*, 26.

VERBA, S. AND N. H. NIE (1971), *Participation in America: Politcal Democracy and Social Equality*, New York: Harper and Row.

VUONG, Q.H. (1985), "Two-Stage Conditional Maximum Likelihood Estimation of Econometric Models," manuscript, California Institute of Technology.

WOLFINGER, R. E. and S. J. ROSENSTONE (1980), *Who Votes?* New Haven, CT: Yale University Press.